LIBRARY OF THE
UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

016.7114
C73e
no.495-509

016.7 ref. cpl
C73e
#495-509

UNIVERSITY OF ILLINOIS LIBRARY

Name

Identification Number 1-31-74

Address

VIRA SACHAKUL
NAT'L CLEARINGHOUSE
DEPT. OF ARCH.
[illegible]

Digitized by the Internet Archive
in 2010 with funding from
University of Illinois Urbana-Champaign

http://www.archive.org/details/twomodesofinform500gran

Council of Planning Librarians EXCHANGE BIBLIOGRAPHY

December 1973

500

TWO MODES OF INFORMATION RETRIEVAL: A Compact, Bibliographic Distinction Between The Question Answering Systems And Literature Searching Systems

Donald P. Grant, Architect and Planner
Associate Professor, Cal. Poly., San Luis Obispo
Editor, DMG-DRS Journal; Design Research Methods
Chairman, The Design Methods Group

The Library
DEC 3 1973
University of Illinois
at Urbana-Champaign

Mrs. Mary Vance, Editor
Post Office Box 229
Monticello, Illinois 61856

TWO MODES OF INFORMATION RETRIEVAL:

A COMPACT, BIBLIOGRAPHIC DISTINCTION BETWEEN THE QUESTION ANSWERING SYSTEMS AND LITERATURE SEARCHING SYSTEMS

by

Donald P. Grant, Architect and Planner
Associate Professor, Cal Poly, San Luis Obispo
Editor, DMG-DRS JOURNAL: DESIGN RESEARCH AND METHODS
Chairman, THE DESIGN METHODS GROUP

Research and study in producing this bibliographic paper were supported by a grant from the Graham Foundation for Advanced Studies in the Fine Arts, for work towards a Ph.D. in Architecture at the University of California at Berkeley. Minor emphasis work in Information System Design is the basis for this survey of two approaches to information retrieval.

In his characterization of information systems for planning and design, Horst Rittel distinguishes data from information. Data is in the form of documents; information is a process. One way of viewing the distinction is to say that information results from the interpretation of data within one's worldview. For a process of information to have taken place, there must have been a change in knowledge: contradiction, corroboration, alteration, displacement; but a change.

In the field of information system design, a distinction is drawn between question answering systems and literature searching systems. One question to be dealt with in the design of a planning and design information system is the choice between these two morphologies; there are, of course, many additional issues to be resolved that go beyond this distinction.

In this bibliographic paper, the two modes of information retrieval, the question answering system and the literature searching system, are defined, and the literature dealing with each is surveyed. It is suggested that the most useful approach for use in planning and design information systems is the literature searching (i.e.; reference retrieval systems, document retrieval systems) approach, since it is more "friendly" to the interpretation of data within a user's worldview than the more self-contained question-answering approach. Two approaches to information retrieval are the question answering system and the literature searching system. Question answering systems, also known as fact retrieval systems (Cooper, 1964), information retrieval

systems (Bar-Hillel, 1957) and data-providing systems (Bar-Hillel, 1960), provide direct answers from stored knowledge to questions of the form, "What is the answer to ...?" Literature searching systems, also known as reference-providing systems (Bar-Hillel, 1960) or document retrieval systems (Cooper, 1964) provide documents or citations in response to questions of the form, "What literature pertains to ...?" Both question answering and literature searching systems attempt to analyze and identify the information that is necessary to meeting an information need, and only that which is necessary, and provide it to the user. Information can be taken as that which will change one's cognitive map; an information need exists when one encounters a gap in that map (Maron, 1971-242). Rittel defines information as a process that changes one's knowledge (Rittel, 1971-239). McKay defines information in terms of determing action toward goal states (McKay, 1971). In general, the question answering system attempts to directly provide the data that will inform, while the literature searching system attempts to direct the user to those sources in which such data can be found. The reference retrieval or literature searching system does the first part of the fact retrieval or question answering system's job.

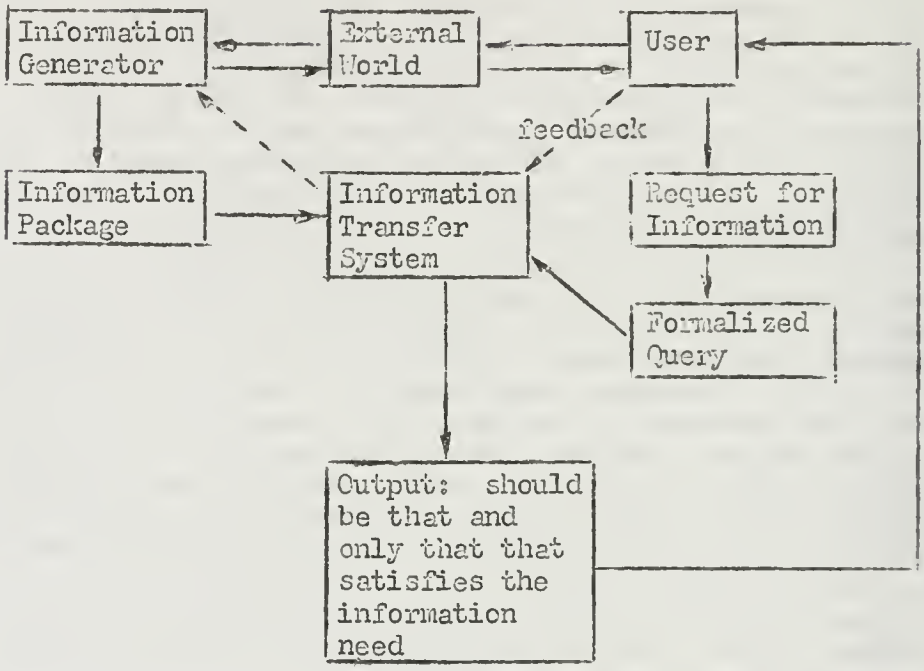
Of conceivable question answering systems, the simplest are the fill-in-the-blank, look-up-device variety, in which the chief interest is in combining the talents of man and machine in order to rapidly access specific data from a large store. The most complex are on the order of essay writing, conversational and picture-describing machines, almost at the level of Turing's challenge to construct a thinking machine. Related areas are artificial intelligence, heuristics for logical inference and automated theorem proving.

Simple and useful question answering systems that deal with queries of a constant form against a rapidly changing store of data are presently in use, for example, the airline reservation system. More complex and difficult systems attempt to carry out logical deductions based on stored data and questions that are input in natural language. One form of question answering system centers around the concept of a logical consequence detector that stores a body of reference sentences and checks test sentences against them in terms of whether or not the test sentences are logical consequences of the stored sentences. Cooper (1964) constructed a question answering system based on the propositions that such a system should accept its stored sentences and its test sentences in the form of natural language and that it should possess the capability of performing logical deductions among these sentences; thus the key effort was toward the development of a system of logical inference among natural language sentences. In this system, both test and stored sentences

are converted from natural language to a logical language devised for the system, then compared. If a subset of the stored sentences is found such that the test sentence is a logical consequence of the subset, the system answers "true" to the test sentence; if not, and a further search for subsets yields one that is logically inconsistent with the test sentence, the system answers "false". If neither logical consequence nor logical inconsistency can be found, the system answers "don't know". The body of stored sentences was very small, yet the number of manipulations that was necessary in seeking subsets was very large. Further progress would seem to rest on a deeper and more precise understanding of logic and language, and some means of dealing with the extremely large numbers resulting from the manipulation of a large body of stored sentences. Simmons (1967), in a survey of the first five years of research into question answering systems, concludes that the fundamental questions in the area are those of how to characterize sentence meaning, how to deal with semantic and/or syntactic ambiguity, how to make inferences without an exhaustive field coverage of all combinatorial possibilities, how to combine partial answers from widely separated points in the data body, to what extent can or should natural language be translated into formal language for manipulation, and whether further work in question answering systems should be basically theoretical or empirical based on large question answerers built for test purposes.

The literature searching system accepts queries of the form, "What literature pertains to ...?", and outputs bibliographies, stacks of documents, or in some cases ranked lists of probability ordered documents. The emphasis in the design of such systems is on their statistical performance over time. Central topics are conceptual closeness, relevance, associative search, probabilistic indexing, automated text analysis and indexing, measures of retrieval effectiveness, clustering, and user inference searching.

The functioning of an information retrieval system can be diagrammed as follows:



The search models by which the information transfer system might proceed can be characterized as direct match, simple overlap, Boolean overlap and feedback models. Assuming a collection of documents, each of which is assigned a series of index terms, and formal queries composed of series of index terms:

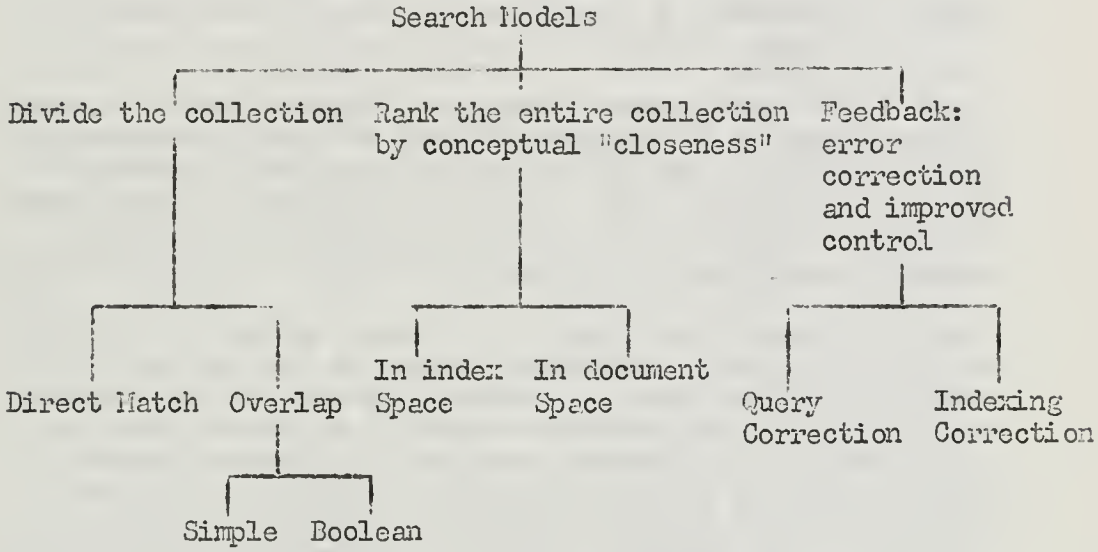
1. In the direct match mode, only those documents will be accessed that have precisely the same list of index terms as are in the formal query.

2. In simple overlap, those documents will be accessed that have some specified minimum number of coinciding index terms.

3. In Boolean overlap, the matching will be based on strings of index terms connected by "and", "or", and "not", so that accessed documents will have to meet certain specified combinations of possible terms in the Boolean query.

4. In the feedback model, the user will be informed of the number of documents that would be accessed in response to his query, and allowed to loosen or tighten his specification in order to increase or decrease the number of documents accessed; and then to modify his query in response to what is accessed.

The mode of system response can further be characterized as either dividing the document collection into two parts, one of which is predicted to be relevant and the other not; or ranking the entire document collection in terms of the probability of each document's satisfying the information need.



The problem of conceptual closeness can be approached as being one of statistical correlation. Kuhns (1964) proposes a series of measures of association or closeness in index space in terms of a statistical measure of independence. If in terms of the co-occurrence of two index terms, A and B, on documents, the probability of the occurrence of A given B is the same as the probability of A, then A and B are said to be independent. Kuhns derives a statistical measure of independence. For a given body of documents, a statistical measure of co-occurrence can be derived that can be compared with the measure of independence. Each pair of terms A and B can be said to be more or less associated depending on the excess of their statistical measure over the independence value. The measure of association is computed for all pairs of terms for an entire collection. Then, for a given search, the number of documents can be expanded by expanding the query to include terms highly associated with those that make up the initial query. This expansion by means of associative searching can be approached using a chain strategy, wherein the search is carried out for the original term, then the term most highly associated with the original term, then the term most highly associated with the first associated term, and so on, or a hub strategy, wherein the search is expanded first with the term most highly correlated

with the original term, then the second most highly associated term with the original term, and so on, or by some combination of these two strategies. Kuhns claims that all "reasonable" measures of association are of the form:

$$\frac{\text{Excess over independence value of A and B}}{\text{A function of the four categories } u, v, x, \text{ and } y,}$$

where: x = A and B; y = not A and not B; u = A and not B; v = not A and B.

Cooper (1971-241) proposes a measure of association in document space obtained by computing the degree of association between pairs of documents for the classes of documents indexed by two different index terms. One of the measures of association proposed by Kuhns can then be applied to get a document to document correlation for all pairs of documents in the collection.

Maron and Kuhns (1960) propose a technique called "probabilistic indexing" wherein a comparative concept of relevance is developed based on theory of probability and used for indexing and searching in an automated document retrieval system. In this technique, the computer computes a "relevance" number for each document in the collection in response to a given request, and then ranks the entire collection by probable relevance. This approach is one of statistical closeness rather than semantic closeness; it interprets a request as a clue on which to base statistical inference.

Closeness and the concepts of associative search and probabilistic indexing are attempts to deal with the problem of relevance. A system should access those documents in response to a given request that are relevant; i.e., it should attempt to maximize the number of documents that are relevant-and-retrieved and minimize the number of documents that are not-relevant-and-retrieved and that are relevant-and-not-retrieved. Common measures of the effectiveness of a system in retrieval are those of precision (avoidance of retrieving irrelevant documents) and recall (avoidance of not retrieving relevant documents). Cooper (1968) proposes a measure of retrievel effectiveness based on expected search length, based on the expected number of irrelevant documents in a collection that would have to be searched through before a specified desired number of relevant documents could be found. The advantages of Cooper's "expected search length reduction factor" over the traditional measures are in providing a single measure of performance, gradations of retrieval status, evaluation relative to random searching, and the taking into account of the amount of relevant material desired by the user. Cooper (1970) goes further to explore the relations between expected search length as a performance measure and system characteristics of three types of retrieval systems: a single-term-request,

direct match system, a multiple-term-request, simple overlap system, and what he describes as a "quasi-Boolean" overlap system that allows the prefixing of the one or more terms in the request by "not". These case studies were undertaken to explore the effect of various system characteristics on retrieval effectiveness.

Relevance can be viewed as being either logical in nature, based on the topic itself, and relations of "appropriateness", "pertinence", or "direct answer-ness" between the query and a document; or utilitarian, based on recency, quality, accuracy, language in which written, and authorship. Maron and Kuhns (1960) in their paper on Probabilistic Indexing develop a probability-based comparative concept of relevance that incorporates statistical measures of user satisfaction with documents in a collection. Relevance in this view is roughly the probability of satisfying an information need. Cooper (1969) describes a precise definition of relevance based on his logical consequence detecting question answering system (1964): that is, that a stored statement is relevant to a query if and only if it is a member of a minimal premise set of which the query is a logical consequence for either the query or its negation. This is a mathematically precise definition only for certain inferential data retrieval systems, but can be conceptually extended to reference retrieval systems: "A document is relevant to an information need if and only if it contains at least one sentence which is relevant to that need." (Cooper, 1969, p. 10). Cooper concludes by contrasting utility and relevance in measuring the performance of retrieval systems and states that perhaps the relevance of "relevance" is that it is the most important factor in utility (1969, pp. 21-22). An important factor in the utility of a specific document to a specific request might be its dependency on previously read documents; document D2 might be useful only after D1 has been read, or might be useful only if D1 has not been read, and so on. Kuhns describes a measure of performance related to the willingness of the user to process more than a certain number of documents before reaching a minimum number of relevant documents (unpubl. NASA-TRW report; Maron, 1971-242).

Useful measures of retrieval effectiveness should be comparative, should be quantitative, and perhaps should carry some concept of the cost of using the system. Swanson (1960) incorporates a penalty for irrelevance into the measure of effectiveness. Swets (1963, 1969) places emphasis on the need for the measure of performance to be a single number, rather than a set of numbers such as precision and recall measures. He describes a probabilistic approach to optimizing system performance.

Salton (1969) compares conventional, manual document indexing by human indexers with fully automatic text processing, and concludes that fully automatic text analysis and search systems do not appear to perform in a manner inferior to manual document indexing and search formulation, and that while manual systems can range from very excellent to very poor, depending on operator awareness of user needs and the stored collection, automated systems perform more or less consistently. Automatic indexing can be approached by simply throwing out all connectives and non-content words and listing the remaining substance words, and counting the number of times that they occur. The result would be a list of index terms ranked by their frequency of occurrence in the document. Given search queries with enough clues (terms) and such modifications as weighted index terms, an automated indexing and searching system should be able to do very well in comparison to manual indexing and searching systems. Kuhn distinguishes between derivative indexing (using terms derived from the text) and assignment indexing (using terms from a standard list)(1962), and formulates the problem as being to disregard syntactic function words and infrequently occurring words so that the remaining high frequency words might give a reasonable picture of "aboutness". Among several objections to this definition (Stevens, 1970, p. 173) is the one that the most information might well be conveyed by the least frequently occurring terms. Stevens describes a large quantity of research into the area of automated indexing, and concludes that the field is already at the stage where usable applications can be made, and foresees expansion of the field in order to gain the benefits of more detailed and broader indexing with given resources and easy updating and revision of indexing.

I. J. Good cites several fields wherein the concept of clustering is being explored. A cluster is simply a group of similar individuals, which in information retrieval might be either index terms or documents. Searching in a clustered collection would be in cluster sets. Libraries have traditionally been divided into subjects; perhaps it would be beneficial to make the division in terms of similar (clustered) vectors of descriptors. Clustering for information retrieval would be superfluous if infinite search speeds were available, but since they are not, it may prove to be a useful technique. Clustering techniques are described by Ball (1966), Sokal (1966), Bonner (1964) and MacNaughton-Smith (1965).

Another approach to searching might be formulated based on the user's making inferences about the documents that he thinks will satisfy his information need, based on qualities that might accompany the desired traits. In this approach, the user might put together a series of contextual characteristics, such as "people who have co-authored with....," "publish in

the journal of....", and so on. To this end a relational language, or calculus of relations might prove useful.

Cooper (1969, p. 14) characterizes the difference between question answering (inferential) systems and literature searching (reference retrieval) systems as follows: "In the inferential system, the only information used in deducing an answer is information stored within the system memory, and all of the logical deductions are also carried out by the system itself. In a reference retrieval system, the user's own information in his personal memory is also brought into play, and the user contributes his own powers of reasoning to the deductive process as well. Thus the stored information available for satisfying an information need with the help of a reference retrieval system must be viewed as the user's own knowledge plus the system stored data, and the deductive apparatus is the user's own deductive power plus whatever deductive power has been programmed into the system."

Note that the manner in which Cooper defines the use of a reference retrieval system closely parallels characterizations of information in the planning process: one's values or worldview, combined with factual data, to yield plans for action in bringing about a desired situation. One might conclude from this, that the greatest usefulness of formal information retrieval systems in planning and design will be in the use of reference retrieval systems to provide factual data for interpretation by the user's deontic (ought-to-be) knowledge; and that the uses of question answering systems will probably be on only the simpler levels, parallel to airline reservation systems, in which the emphasis is on putting standard queries to constantly changing data bases.

Maron (1971-242) describes the field of information retrieval as being on a plateau at present. A large quantity of work has been done over the last 10-15 years, but now there seems to be a lack of progress. Perhaps what is lacking is a theoretical basis for the next breakthroughs, and that basis must come in some area such as the study of information transfer in the brain. One area of great interest in the near future in the field of information retrieval is likely to be that of undesirable side effects such as authoritarian misuse of data on individuals, accessible and manipulable in a way that has never before been the case. There is current interest in the technological forecasting of such abuses and how they might be controlled.

Ball, G. H. (1966) A COMPARISON OF SOME CLUSTER SEEKING TECHNIQUES. SRI Tech. Rept. RADC-TR66-514, November 1966.

Bar-Hillel, Y. (1957) "A Logician's Reaction to Recent Theorizing on Information Search Systems." AMERICAN DOCUMENTATION. Vol. 8, No. 2, April 1957.

Bar-Hillel, Y. (1960) "Some Theoretical Aspects of the Mechanization of Literature Searching." TECHNICAL REPORT NO. 3. Contract no. N62558-2214, Off.Techn. Ser., US Dept. of Commerce. Washington, D.C., April 1960.

Bonner, R. E. (1964) "On Some Clustering Techniques." IBM JOURNAL OF RESEARCH AND DEVELOPMENT. Vol. 8, No. 1, January 1964, 22-32.

Cooper, W. S. (1964) "Fact Retrieval and Deductive Question-Answering Information Retrieval Systems." JOURNAL OF THE ASSOCIATION FOR COMPUTING MACHINERY. Vol. 11, No. 2, April 1964, 117-137.

Cooper, W. S. (1968) "Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems." AMERICAN DOCUMENTATION. Vol. 19, No. 1, January 1968, 30-41.

Cooper, W. S. (1969) A DEFINITION OF RELEVANCE FOR INFORMATION RETRIEVAL. Institute of Library Research, University of California, Technical Paper Series, Paper No. 10, December 12, 1969, Berkeley.

Cooper, W. S. (1970) "On Deriving Design Equations for Information Retrieval Systems." JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE. November-December 1970, 385-395.

Cooper, W. S. (1971-241) Lecture in Librarianship 241, Winter 1971, University of California, Berkeley. Course title: INFORMATION TRANSFER AND RETRIEVAL.

Good, I. J. (1962) "Botryological Speculations," in Good, I. J. A SCIENTIST SPECULATES: AN ANTHOLOGY OF PARTLY BAKED IDEAS. Basic Books, 120-132.

Kuhns, J. L. (1964) "The Continuum of Coefficients of Association." SYMPOSIUM PROCEEDINGS: STATISTICAL ASSOCIATIVE METHODS FOR MECHANIZED DOCUMENTATION. National Bureau of Standards Misc. Rept. No. 269, 33-39.

Luhn, H. P. (1962) "Automated Intelligence Systems," in Hattery, L. H. and McCormick, E. M. INFORMATION RETRIEVAL MANAGEMENT. Detroit: Am. Data Proc., Inc.

MacNaughton-Smith, P. (1965) SOME STATISTICAL AND OTHER NUMERICAL TECHNIQUES FOR CLASSIFYING INDIVIDUALS. London: H.M.S.O.

McKay, D. C. (1971) Lecture at the School of Librarianship. University of California, Berkeley, April 13, 1971.

Maron, M. E. and J. L. Kuhns. (1960) "On Relevance, Probabilistic Indexing and Information Retrieval." JOURNAL OF THE ASSOCIATION FOR COMPUTING MACHINERY. Vol. 7, No. 3, July 1960, 216-244.

Maron, M. E. (1971-242) Lectures in Librarianship 242A and 242B, Winter and Spring 1971, University of California, Berkeley. Course title: FORMAL TECHNIQUES FOR INTELLECTUAL ACCESS.

Rittel, H. W. J. (1971, 239A) Lectures in Architecture 239A, Spring 1971, University of California, Berkeley. Course title: INFORMATION SYSTEMS FOR PLANNING AND DESIGN.

Salton, G. (1969) "A Comparison Between Manual and Automatic Indexing Methods." AMERICAN DOCUMENTATION. Vol. 20, No. 1, January 1969, 61-71.

Simmons, R. F. (1967) "Answering English Questions by Computer," in Borko, H. AUTOMATED LANGUAGE PROCESSING. New York: Wiley.

Sokal, R. R. (1966) "Numerical Taxonomy." SCIENTIFIC AMERICAN. December 1966.

Stevens, Mary E. (1970) AUTOMATIC INDEXING: A STATE OF THE ART REPORT. Washington, National Bureau of Standards, U. S. Government Printing Office.

Swanson, D. (1960) "Searching Natural Language Text by Computer." SCIENCE. Vol. 132, No. 3434, October 21, 1960, 1099-1104.

Swets, J. A. (1963) "Information Retrieval Systems." SCIENCE. Vol. 141, No. 3577, July 19, 1963, 245-250.

Swets, J. A. (1969) "Effectiveness of Information Retrieval Methods." AMERICAN DOCUMENTATION. Vol. 20, No. 1, January 1969, 72-89.

COUNCIL OF PLANNING LIBRARIANS Exchange Bibliography #500

TWO MODES OF INFORMATION RETRIEVAL:

A COMPACT, BIBLIOGRAPHIC DISTINCTION BETWEEN THE QUESTION ANSWERING SYSTEMS AND LITERATURE SEARCHING SYSTEMS

Additional copies available from:

Council of Planning Librarians
Post Office Box 229
Monticello, Illinois 61856

for $1.50

BOUND BY
HNM

UNIVERSITY OF ILLINOIS-URBANA
016.7114C73E C001
EXCHANGE BIBLIOGRAPHY. URBANA, ILL
495-509 1973
3 0112 029109268